

REMARKS

Claims 1-16 and 35-50 are pending. Claims 35-50 are new. Claims 17-34 have been canceled in response to the Office Action's request for clarification of the numbering. Claims 19-34 presented in the Response filed March 18, 2005 now correspond to new claims 35-50, respectively. Likewise, claims 17-32 presented in the Response filed November 7, 2005 also correspond to new claims 35-50, respectively.

Claims 1 and 2 are herein amended. Claims 3-9, 11 and 16 were previously amended.

Claims 35 and 36, while new, are modified from now canceled claims 17 and 18 as filed in the Response of November 7, 2005 in the same manner as claims 1 and 2 have been amended. Support for these amendments and modifications is fully described below in response to the Office's request to amend respective terms in each of the foregoing claims.

New claims 37-50 are copied directly from claims 19-32 as presented in the Response filed November 7, 2005.

For the Examiner's convenience, responses herein have been numbered to correspond to the appropriate objection or rejection in the Office Action.

It should be noted that the correct Attorney Docket No. for this application is "77670/593."

New Matter Objection to Amendment to Specification

[6] The Office Action objects to amendments to the specification filed on March 18, 2005. The Office asserts that Applicants have failed to point to express support for the recited limitation $D_1D_2X_1(X_2X_3)X_4D_3$. The Office suggests Applicants have argued that support for $D_1D_2X_1(X_2X_3)X_4D_3$ is only *implicitly* supported by specific examples in the specification and these specifically disclosed species fail to describe the genus of sequences having the formula $D_1D_2X_1(X_2X_3)X_4D_3$. The Office further urges that Applicants do not dispute $D_1D_2X_1(X_2X_3)X_4D_3$ alters the positioning of the parentheses such that it encompasses amino acid sequence permutations that are not supported by the original disclosure.

Applicants respectfully traverse the Office's assertions and conclusions. First, the expressly disclosed $D_1D_2X_1X_2(X_3X_4)D_3$ is mathematically equivalent to $D_1D_2X_1(X_2X_3)X_4D_3$ and the genus of sequences encompassed by these two algorithms is exactly the same genus. Second,

in the response of November 7, 2005, Applicants provided a mathematical proof that $D_1D_2X_1X_2(X_3X_4)D_3$ encompasses the same sequences as $D_1D_2X_1(X_2X_3)X_4D_3$. As such, Applicants again traverse the assertion that $D_1D_2X_1(X_2X_3)X_4D_3$ creates sequence permutations that are not supported by the original disclosure.

Contrary to the Office's understanding, support for $D_1D_2X_1(X_2X_3)X_4D_3$ is not *implicit* in $D_1D_2X_1X_2(X_3X_4)D_3$, instead, support is explicit because the algorithm $D_1D_2X_1X_2(X_3X_4)D_3$ is mathematically exactly equivalent to $D_1D_2X_1(X_2X_3)X_4D_3$ and, as such, these two algorithms encompass the exact same genus of sequences. Applicants provide the following chart to demonstrate the mathematical equivalency of $D_1D_2X_1(X_2X_3)X_4D_3$ and $D_1D_2X_1X_2(X_3X_4)D_3$ and the fact that these algorithms encompass the exact same genus of sequences.

D₁D₂X₁X₂(X₃X₄)D₃

Possible outcomes where
 X is any amino acid and
 X₃ and/or X₄ are present or absent

I. D₁D₂X₁X₂D₃
 X₃ and X₄ are absent

II. D₁D₂X₁X₂X₃D₃
 X₄ is absent

III. D₁D₂X₁X₂X₄D₃
 X₃ is absent

IV. D₁D₂X₁X₂X₃X₄D
 X₃ and X₄ are present

D₁D₂X₁(X₂X₃)X₄D₃

Possible outcomes where
 X is any amino acid and
 X₂ and/or X₃ are present or absent

I. D₁D₂X₁X₄D₃
 X₂ and X₃ are absent

Since X₄ is any amino acid, the range of possible amino acids for X₄ is the same as the range of possible amino acids for X₂ in the left hand column:
 $X_4^{(\text{amino acid 1-20})} = X_2^{(\text{amino acid 1-20})}$.
 D₁D₂X₁X₄D₃, therefore, represents the same range of sequence possibilities as D₁D₂X₁X₂D₃ (Section I) in the left column.

II. D₁D₂X₁X₂X₄D₃
 X₃ is absent

Since X₄ is any amino acid, the range of possible amino acids for X₄ is the same as the range of possible amino acids for X₃ in the left hand column:
 $X_4^{(\text{amino acid 1-20})} = X_3^{(\text{amino acid 1-20})}$.
 D₁D₂X₁X₂X₄D₃, therefore, represents the same range of sequence possibilities as D₁D₂X₁X₂X₃D₃ (Section II) in the left column.

III. D₁D₂X₁X₃X₄D₃
 X₂ is absent

Since X₃ is any amino acid, the range of possible amino acids for X₃ is the same as the range of possible amino acids for X₂ in the left hand column:
 $X_3^{(\text{amino acid 1-20})} = X_2^{(\text{amino acid 1-20})}$,
 D₁D₂X₁X₃X₄D₃, therefore, represents the same range of sequence possibilities as D₁D₂X₁X₂X₄D₃ (Section III) in the left column.

IV. D₁D₂X₁X₂X₃X₄D₃
 X₂ and X₃ are present

D₁D₂X₁X₂X₃X₄D₃ are equivalent to D₁D₂X₁X₂X₃X₄D₃ in (IV) in the left column

The foregoing table demonstrates that, contrary to the Office's conclusion, the position of the parenthesis in the algorithm creates no change in the genus of amino acid sequences encompassed therein. Hence, Applicants submit that the genus of the sequence algorithm $D_1D_2X_1(X_2X_3)X_4D_3$ is fully supported in the specification.

Furthermore, as extensively set forth in the response filed November 7, 2005, the application is replete with sequences that reflect the algorithm $D_1D_2X_1(X_2X_3)X_4D_3$. *See, e.g.*, Figure 1 and column 4, lines 60-67; column 6, lines 48-63. Hence, the sequence $D_1D_2X_1(X_2X_3)X_4D_3$ in region II of the prenyl diphosphate synthases is supported by the original disclosure at, for example, column 4, line 24. Withdrawal of the objection is therefore respectfully requested.

Objection to format of amended claims

[7] The Office Action objects to claims 17-32 for improper format. In response to the Office's objection, claims 17-34 have been canceled. Canceled claims 19-34 as presented in the Response filed March 18, 2005 now correspond to new claims 35-50, respectively. Likewise, canceled claims 17-32, as erroneously presented in the Response filed November 7, 2005, also correspond to new claims 35-50, respectively.

Claims 35 and 36, while new, are modified from now-canceled claims 17 and 18 as filed in the Response of November 7, 2005. The modifications contained in claims 35 and 36 correspond to the amendments to claims 1 and 2. Support for these amendments and modifications is fully described below in response to the Office's request to amend respective terms in each of the foregoing claims.

New claims 37-50 are copied directly from claims 19-32 as presented in the Response filed November 7, 2005.

Because new claims 35-50 comply with 37 C.F.R. § 1.173(b)(2), Applicants respectfully request the Office withdraw its objection for the format of the claims.

Office request and support for amendment to claims 1 and 17 to clarify synthesis of shorter chain lengths

[8] The Office Action requests that claims 1 and 17 be amended to clarify how prenyl diphosphate synthesized by a mutant prenyl diphosphate synthase may be shorter than that

synthesized by a corresponding wild-type enzyme. Claim 1 and claim 35 (corresponding to claim 17) have been amended/modified to make such a clarification. The amendment to claim 1 and modification of new claim 35 clarify that the mutant enzyme synthesizes a greater amount of prenyl diphosphate of a shorter length than does a wild-type enzyme. This amendment finds support, for example, in Figure 3; the illustration of Figure 3 at column 5, lines 15 to 28; and throughout the specification, including column 6, lines 17 to 21. As such, Applicants respectfully request the Office withdraw the rejection of claim 1 for clarity of synthesis of shorter chain lengths and enter new claim 35 for the same reason.

Office request and support for amendment to claims 1 and 17 to clarify "Region II"

[9] The Office has requested in the Office Actions of July 18, 2005 and January 3, 2006 that the claims be amended to clarify the position of the acid-rich domain of Region II by identifying distinguishing characteristics of Region II. In response to the Office's request, the distinguishing characteristics of Region II are presently clarified in amended claim 1 and new claim 35 by establishing that the acid-rich domain sequence $D_1D_2X_1(X_2X_3)X_4D_3$ or $D_1D_2X_1X_2(X_3X_4)D_3$ (which are nevertheless identical) must be contained within region II and the sequence of region II must share at least 25% homology with region II of SEQ ID NO:2. SEQ ID NO:2 is a geranylgeranyl diphosphate synthase of *Arabidopsis thaliana* that is disclosed at column 4, line 60 through 64 and in Figure 1.

Support for the amendment to claim 1 and for new claim 35, which clarify the term "Region II," may be found, for example, in claim 1, as originally filed; Figure 1 and the illustration of Figure 1 at column 4, line 60 through column 5, line 7; Example 1, in column 10; Example 4, in column 12; Chen *et al.*, Protein Science Vol. 3, pp. 600-607 (1994) (cited and discussed in application at column 5, lines 31-63) (a copy of which is provided herewith); and in the art as discussed by K. Kelly, "Exhaustive and Iterative Clustering of the Protein Databank" JCCG Winter 1998, <http://www.chemcomp.com/journal/families.htm> (a copy of which is provided herewith).

Claim 1, as originally filed, recited the aspartic acid-rich domain of Region II of prenyl diphosphate synthase. As such, the application as filed supports this particular characterization of Region II.

Support for the percent homology limitation of the claims, as amended, is found in all of the prenyl diphosphate synthase sequences disclosed in the application. It is further supported by knowledge in the art at the time of filing. One of skill in the art at the time of filing fully understood that Region II of prenyl diphosphate synthases could be determined by locating the aspartic acid rich region $D_1D_2X_1(X_2X_3)X_4D_3$ and implementing pairwise comparisons to determine the homology of Region II as compared to a known prenyl diphosphate synthase region II. The skilled artisan's understanding is fully disclosed in Chen *et al.*, Protein Science Vol. 3, pp. 600-607 (1994), the findings of which are discussed in the application at column 5, lines 31 through 63.

Chen demonstrates that the art had fully identified homologous Region II in all prenyl diphosphate synthases sequenced (13 total synthase sequences). Chen further demonstrates that Region II of each of the thirteen sequences shares at least 25% homology with Region II of SEQ ID NO:2. *See, e.g.*, page 602, Figure 2, Region II (compare GGPP_CAN to all other sequences). Furthermore, in the application as filed, all sequences denoted as Region II of prenyl diphosphate synthases share at least 25% homology with positions 72 to 93 (Region II) of SEQ ID NO:2. *See, e.g.*, sequences of Figure 1; mutant nucleic acid sequences of SacGGPS in Example 4 (col. 12, lines 16-43); and wild-type sequence of SacGGPS (GenBank accession no. D28748) in Example 1 (col. 10, lines 20-26).

Further support for the amendment may be found by reference to what was well known in the art at the time the application was filed. Applicants have attached hereto K. Kelly, "Exhaustive and Iterative Clustering of the Protein Databank" JCCG Winter 1998. Kelly references the state of the art in early 1998 when acknowledging:

To date, the most useful principle which is applied to guide such searches is the rule that 'similar sequence' implies 'similar structure' Generally speaking, if a protein sequence shares more than 25% pairwise similarity with a known structure, it usually also shares at least the broad outlines of the fold topology of the known structure.

See id. at 1 (emphasis added). In view of the teachings of Kelly and Chen, one of skill in the art would expect a sequence from a prenyl diphosphate synthase comprising $D_1D_2X_1(X_2X_3)X_4D_3$ and having at least 25% homology with Region II of SEQ ID NO:2 to possess the topological

and functional characteristics of Region II as taught by Chen. *See* Chen at 605, Fig. 6 (disclosing Region II of prenyl diphosphate synthases as juncture of α -helix, $\alpha 2$, and loop, L3).

In view of the homology of the sequences disclosed in the application, including the sequences disclosed in Chen, and the understanding in the art of topological and functional characteristics of Region II, as taught in Chen and Kelly, Applicants respectfully submit the characterization of Region II of prenyl diphosphate synthases in amended claim 1 and new claim 35, as urged by the Examiner, is fully supported by the application as filed. As such, Applicants respectfully request the Office withdraw the rejection of claim 1 for clarity of Region II of prenyl diphosphate synthase and likewise enter new claim 35 containing the same clarification.

Office request and support for amendment to claims 2, 16 and 18 to clarify “enzymatic activity”

[10] The Office Actions requests amendment to claims 2, 16 and 18 (new claim 36) to clarify the term “enzymatic activity.” Claim 2 and new claim 36 have been amended/modified to replace the term “enzymatic activity” with “synthesizes about as much or more prenyl diphosphate than the amount of prenyl diphosphate synthesized by the wild type prenyl diphosphate synthase under similar conditions.” Support for this amendment may be found, for example, in Figure 2; the illustration of Figure 2 at column 5, lines 8 to 15; and in Example 5 at column 12, line 45 through column 13, line 14. This disclosure shows that the enzymatic activity of the synthases was determined by counting total incorporation of radioactive isopentenyl diphosphate into prenyl diphosphates generally. As such, enzymatic activity reflects a measure of the total synthesis of the broad class of prenyl diphosphates by prenyl diphosphate synthases. Claim 2 and new claim 36, as amended, reflect this clarification of the term “enzymatic activity.” As such, Applicants respectfully request the Office withdraw the rejection of claim 2 for clarity of “enzymatic action” and likewise enter new claim 36. Likewise the rejection of claim 16, which depends from claim 2, should be obviated.

Rejection of claims 17-32 under 35 U.S.C. § 112, first paragraph

[11] The Office Action rejects claims 17 to 32 (new claims 35-50) for lack of written description because the Office believes the genus of sequences defined by $D_1D_2X_1(X_2X_3)X_4D_3$ is different from the genus of species defined by $D_1D_2X_1X_2(X_3X_4)D_3$. However, as Applicants have explained above, $D_1D_2X_1(X_2X_3)X_4D_3$ and $D_1D_2X_1X_2(X_3X_4)D_3$ encompass the very same

genus. Because the genus of $D_1D_2X_1(X_2X_3)X_4D_3$ is the same as the genus of $D_1D_2X_1X_2(X_3X_4)D_3$, Applicants respectfully submit the presence of $D_1D_2X_1(X_2X_3)X_4D_3$ in claims 35 to 50 is not new matter and respectfully request withdrawal of the rejection of claims 35 to 50 for lack of support.

Rejection of claims 7, 16 and 23 under 35 U.S.C. § 112, first paragraph

[12] The Office Action rejects claims 7 and 23 (new claim 41) for lack of written description and invites Applicants to show support for the limitation of thermostability greater than wild-type. Support for claim 7 and new claim 41 may be found, for example, in Figure 2 and the illustration of Figure 2 at column 5, lines 7 to 14, where Applicants disclose mutant enzymes having greater relative enzymatic activity than wild-type enzyme as the temperature of the environment is increased. The maintenance of enzymatic activity of the mutants relative to two wild-type enzymes as the environment temperature is increased demonstrates the thermal stability of the mutant enzymes. Because the application as filed supports claim 7 and new claim 41, Applicants respectfully request the Office withdraw the rejection of these claims.

Rejection of claims 1-7, 10, 15-23, 26 and 31-32 under 35 U.S.C. § 112, first paragraph

[13] The Office Action rejects claims 1-7, 10, 15-23, 26 and 31-32 under 35 U.S.C. § 112, first paragraph, for failure to provide sufficient structure to demonstrate possession of the claimed subject matter. The Office rejects Applicants' arguments in support of written description of the claims because the Office believes that outside of the claims' minimal recited structural features of region II, the remainder of the structures of the members encompassed by the genus of the claims are completely undefined. The Office concludes that the minimal sequence structure recited in the claims would not impart farnesyl diphosphate synthase activity. According to the Office Action, the recited enzymatic activity may only be achieved by adding amino acids back to the terminal ends of the minimal recited structural feature to reconstruct a properly folded enzymatically active synthase enzyme. Applicants respectfully traverse the Office's conclusions and assertions in turn.

The Office asserts the pending claims are completely undefined outside of the minimal structure recited in the claims. Applicants strenuously disagree with and respectfully traverse this assertion. The claims under examination are highly structured and the scope of the claims is strictly delimited, well beyond the minimal sequence limitations provided in the claims.

First, the claims are directed to an enzyme, not the polypeptide sequence of an enzyme. As such, the enzyme must function as a prenyl diphosphate synthase. Applicants have provided at least 16 structural examples of functional prenyl diphosphate synthases. Furthermore, Applicants cite to art from 1994 that provides at least 13 structural examples of functional prenyl diphosphate synthase from a wide array of taxonomic phyla and provides at least five other articles between 1993 and 1995 disclosing additional examples. Applicants respectfully submit that both the structure and function of prenyl diphosphate synthase was well defined in the art and fully disclosed in the application. One of skill in the art would understand that applicants possessed all of these enzymes and possessed the means to manipulate these enzymes. As such, possession by Applicants of the structural and functional prenyl diphosphate limitation of the claims should not be in doubt. As a result, Applicants traverse the assertion that amino acids must be added back to the structural limitations of the claims to achieve a functioning synthase enzyme. In fact, the claims begin, in their first limitation, with a fully-structured, functioning synthase enzyme which is only then subject to modification. Nothing is added back. Applicants respectfully request the Office withdraw this line of reasoning in support of the rejection of the claims.

Next, the claims are directed to a mutant enzyme defined by the substitution of certain amino acids at particular positions in region II of the wild-type enzyme. Applicants have provided five examples of mutant enzymes with substitutions at the prescribed positions in the sequence of the wild-type enzyme.

Furthermore, Applicants have provided actual working examples of four of five possible substitutions in accordance with element 1(a) of claims 1 and 35, namely, substitution at the fifth position upstream of D_1 (e.g., mutant enzyme 3; column 6, line 53), substitution at the fourth position upstream of D_1 (e.g., mutant enzyme 1; column 6, line 48), substitution at the second position upstream of D_1 (e.g., mutant enzyme 5; column 6, line 59) and substitution at the first position upstream of D_1 (e.g., mutant enzyme 4; column 6, line 56).

Applicants have likewise provided an actual working example of the only substitution possible in accordance with element 1(b) of claim 1, namely, substitution at the first position upstream of D_3 , and the only substitution possible in accordance with element 1(b) of claim 35, namely, substitution at the first position downstream of D_2 , (e.g., mutant enzyme 5, column 6,

line 59). Applicants have also provided an actual working example of both insertions possible in accordance with element 2 of claims 1 and 35, namely, insertion of at least one additional amino acid between D₃ and the first amino acid upstream of D₃ (claim 1) and insertion of an amino acid between the first amino acid downstream of D₂ and the first amino acid upstream of D₃ (claim 35) (*e.g.*, mutant enzyme 5, column 6, line 59).

In view of the working examples disclosed in the application, Applicants have reduced to practice all iterative positions encompassed within the structure provided in the genus of claims 1 and 35 except one, namely, substitution at position 2 upstream of D₁. As discussed more fully below, one of skill in the art (upon review of the prenyl diphosphate enzymes disclosed in the application and known in the art) would recognize Applicants also possessed substitution of residues at position 2 upstream of D₁.

The Office asserts that one of skill in the art would recognize that the “minimal” sequence information in the claims would not impart the recited enzymatic activity. Of course, Applicants do not dispute this. Applicants respectfully submit, however, that the “minimal” sequence information in the claims is not directed to enzymatic activity of the prenyl diphosphate synthase. Instead, the sequence information is directed to modification of the enzymatic activity of a wild-type prenyl diphosphate synthase to impart synthesis of a relatively greater amount of short chain prenyl diphosphate. As such, one of skill in the art would recognize that the wild-type enzyme (fully structured and functional in nature) is enzymatically active and Applicants have provided the artisan with a mechanism for reducing the chain length of the product that results from said already established enzymatic activity.

Applicants respectfully dispute the Office’s conclusion that the claimed elements do not impart enzymatic activity. In fact, in five of five experiments, Applicants have shown that the “minimal” amino acid structure of the claims, in combination with all of the other elements of the claims (*e.g.*, starting material of an enzymatically active enzyme), provides enzymatic activity as claimed. This conclusion is further supported by the results of Chen, wherein thirteen wild-type (enzymatically fully functional) enzymes have variable amino acids at each and all of the positions specified in the claims. *See* Chen at 602, Figure 2 (*e.g.*, FFP_YSC is substituted at positions 1, 2, 4 and 5 upstream from D₁ and position 1 downstream from D₁ of GGPP_CAN).

As such, one of skill would specifically recognize the enzymatic functionality of wild-type enzymes subject to the strictly circumscribed substitutions set forth in claim 1 and claim 35.

Finally, the enzymatic functionality of the mutant prenyl diphosphate synthase is a limitation of the claims. As such, the scope of the claims is directed only to functional mutant enzymes. Since Applicants have provided working examples of functional mutant enzymes with substitutions and insertions at all but one possible positional iteration, as set forth in the claims, Applicants respectfully submit it is immediately apparent to one of skill in the art that Applicants possessed the invention of the claims.

In view of the highly limited structure and well established function of the claims, Applicants respectfully request the Office withdraw the rejection of the claims for having “completely undefined structure” outside of the aspartic-acid rich region and for providing the artisan with insufficient guidance to generate a functional enzyme. The disclosure of the application renders these conclusions manifestly unsubstantiated.

Rejection of claims 1-32 under 35 U.S.C. § 112, first paragraph, for enablement

[14] The Office Action rejects claims 1-32 under 35 U.S.C. § 112, first paragraph, for failure to provide an enabling disclosure. Applicants respectfully traverse the rejection. As set forth immediately above, Applicants have provided working examples of nearly all positional iterations possible within the claims and as well as disclosure demonstrating the extensive number of prenyl diphosphates known in the art, and the broad functional variability of the conserved sequences, which Applicants have taught as candidates for modification. Such disclosure and skill in the art supports a conclusion that the skilled artisan would have understood the examples and teachings of the application can be practiced on any functional wild-type prenyl diphosphate synthase to produce a mutant enzyme of the claims.

The Office has asked Applicants to provide a discussion of enablement within the Wands factors. The Office insists that the claims are overly broad and encompass a vast number of mutant prenyl diphosphate synthases having the minimal structural features of the claims but also encompassing synthases from transgenic organisms and other limitless origins. Applicants respectfully and strenuously traverse the allegation of over breadth. The claims are not overbroad. In fact, they are highly circumscribed. The claims do not encompass a vast number of synthases and do not encompass all synthases from transgenic organisms. To the contrary, the

claims encompass mutants of a wild-type enzyme where substitutions and insertions have occurred at a total of 7 possible positions. Considering that the claimed enzyme must be derived from a wild-type, must be a mutant, must comprise mutations in a highly circumscribed 7 possible positions, and as claimed must produce shorter prenyl diphosphates than the wild-type, the claims are simply extremely narrow.

The Office insists there is a lack of guidance and working examples in the application because only five working examples have been provided. Applicants respectfully traverse this conclusion. As discussed more fully above, the working examples in the specification happen to represent all but one possible positional iteration within the structural limitations of the claims. Furthermore, the art was replete with examples of functional wild-type enzymes, which the skilled artisan would have compared with the data of the application to understand that the claimed enzymes may be derived from any of the functional wild-type enzymes known in the art. This conclusion would have been easily drawn from the extensive conservation of region II and domain I within the synthases of interest. *See* Chen at 602. As such, the guidance within the application is particularly extensive with respect to the universe of possible iterations in the claims. Additionally, the art provided a full understanding of wild-type prenyl diphosphate synthases at the time of filing. In combination, Applicants respectfully request the Office acknowledge that the application as filed contained more than sufficient guidance and more than extensive working examples so long as each limitation of the claims is fully considered in the enablement analysis.

The Office insists there is a high level of unpredictability in the art and cites Branden *et al.* ("Introduction to Protein Structure," Garland Publishing Inc., New York, 1991) for the proposition that protein engineers have frequently been surprised by the range of effects caused by point mutations that they hoped would change only one specific and simple property in enzymes. The Office further notes Branden's assertion that the artisan in 1991 knew little about the rules of protein stability and the artisan found it difficult to design *de novo* stable proteins with specific functions. Applicants respectfully but emphatically traverse these findings.

First, Kelly (1998) (attached herewith) directly contradicts the teaching of Branden (1991) and demonstrates amazing strides in the artisan's understanding of protein structure and function between 1991 and 1997 when the application was first filed.

Next, while a range of effects may be expected by a point mutation in a non-conserved sequence, a point mutation in a highly conserved sequence resulting in a sequence found in other functional enzymes would likely not give rise to unexpected effects. Since the pending claims are directed to mutations in highly circumscribed portions of a highly conserved sequence for which extensive data is available on functional wild-type enzymes, these precisely directed and expressly limited mutations would not be expected to provide unpredictable results. In fact, the expectation would be the opposite. Applicants respectfully submit a high level of predictability in the art with respect to the practice of the claims.

Finally, experimentation would not be undue. The application provides working examples of six of seven possible positions in which mutations may be made. The artisan need only follow the protocol provided in the application to practice the invention on wild-type prenyl diphosphate synthases with mutations created at the seven possible positions.

The Office insists that it is not routine in the art to screen for all polypeptides having a substantial number of substitutions or modifications as encompassed in the claims. Applicants do not disagree with this assertion. Applicants respectfully submit, however, that screening for a substantial number of substitutions or modifications is not at all necessary to practice the invention. In fact, the only screening necessary is the very screening exemplified in the application—for which five of five screens returned working results. As such, Applicants respectfully submit that the easy screening step in the working examples demonstrates in itself that experimentation would not be undue.

Double Patenting Rejections

[15] As noted by the Office Action, Applicants will attend to the double patenting rejections at such time as claims in the present application are allowed. Applicants gratefully acknowledge the Office's disabuse of Applicants' understanding that the double patenting rejection is provisional.

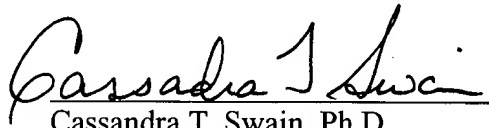
CONCLUSION

The claims are believed to be in condition for allowance and Applicants respectfully request the same. The Examiner is invited to contact the undersigned to discuss any issues related to this application.

The Office is authorized to charge any fees or credit any overpayment regarding this application to Kenyon & Kenyon **Deposit Account No. 11-0600**.

Respectfully submitted,

Date: March 31, 2006


Cassandra T. Swain, Ph.D.
(Reg. No. 48,361)

KENYON & KENYON LLP
1500 K Street, N.W., Suite 700
Washington, DC 20005
Tel: (202) 220-4200
Fax: (202) 220-4201

Isoprenyl diphosphate synthases: Protein sequence comparisons, a phylogenetic tree, and predictions of secondary structure

ANJUN CHEN,¹ PAULO A. KROON,² AND C. DALE POULTER¹

¹ Department of Chemistry, University of Utah, Salt Lake City, Utah 84112

² Department of Biochemistry, University of Queensland, Herston, Queensland 4029, Australia

(RECEIVED November 2, 1993; ACCEPTED February 2, 1994)

Abstract

Isoprenyl diphosphate synthases are ubiquitous enzymes that catalyze the basic chain-elongation reaction in the isoprene biosynthetic pathway. Pairwise sequence comparisons were made for 6 farnesyl diphosphate synthases, 6 geranylgeranyl diphosphate synthases, and a hexaprenyl diphosphate synthase. Five regions with highly conserved residues, two of which contain aspartate-rich DDXX(XX)D motifs found in many prenyltransferases, were identified. A consensus secondary structure for the group, consisting mostly of α -helices, was predicted for the multiply aligned sequences from amino acid compositions, computer assignments of local structure, and hydrophathy indices. Progressive sequence alignments suggest that the 13 isoprenyl diphosphate synthases evolved from a common ancestor into 3 distinct clusters. The most distant separation is between yeast hexaprenyl diphosphate synthetase and the other enzymes. Except for the chromoplastic geranylgeranyl diphosphate synthase from *Capsicum annuum*, the remaining farnesyl and geranylgeranyl diphosphate synthases segregate into prokaryotic/archaeobacterial and eukaryotic families.

Keywords: catalytic site; evolution; farnesyl diphosphate; geranylgeranyl diphosphate; prenyltransferase; secondary structure; substrate binding

With more than 23,000 known members, isoprenoids constitute the most chemically diverse family of naturally occurring compounds. Some of the more important products of the pathway are the sterols (Poulter & Rilling, 1981a), ubiquinones (Ashby & Edwards, 1990), dolichols (Matsuoka et al., 1991), carotenoids (Spurgeon & Porter, 1981), prenylated proteins (Clarke, 1992), and plant mono-, sesqui-, and diterpenes (Cane, 1981; Croteau, 1981; West, 1981). All of these compounds are derived from linear isoprenoid diphosphates synthesized from isopentenyl diphosphate and dimethylallyl diphosphate by a family of prenyltransferases that catalyze sequential condensations of IPP with allylic isoprenoid diphosphates, as shown in Figure 1. Although the chemical mechanisms of these condensation reactions are identical, the isoprenyl diphosphate synthases differ in their selectivity with respect to the chain length and double-bond stereochemistry of their respective allylic substrates and the chain length and stereochemistry of newly formed double bonds in their products (Poulter & Rilling, 1978, 1981b).

During the past few years the structural genes for several farnesyl diphosphate synthases, geranylgeranyl diphosphate synthases, and a hexaprenyl diphosphate synthase have been identified and characterized. Early sequence comparisons revealed 2 conserved DDXX(XX)D aspartate-rich domains (Ashby et al., 1990), which were thought to be binding sites for the diphosphate moieties in IPP and the allylic substrates. This proposal was supported by kinetic studies of site-directed mutants (Marrero et al., 1992). More recently, Koyama et al. (1993) identified 7 conserved regions in eubacterial and eukaryotic FPPSases, including the 2 aspartate-rich regions.

Multiple sequence alignments are valuable for identifying conserved sequences in proteins. In addition, multiple alignments can be used in conjunction with procedures for predicting secondary structure from primary sequences to obtain improved predictions, as for example, the prediction of the structure of the α subunit in tryptophan synthase (Crawford et al., 1987). We now report sequence comparisons for 13 prenyltransferases, including 6 FPPSases, 6 GGPPSases, and a HexPPSase that suggest divergence from a common ancestor based on a com-

Reprint requests to: C. Dale Poulter, Department of Chemistry, University of Utah, Salt Lake City, Utah 84112; e-mail: poulter@chemistry.utah.edu.

Abbreviations: DMAPP, dimethylallyl diphosphate; FPP, farnesyl diphosphate; FPPSase, farnesyl diphosphate synthase; GGPP, geranylgeranyl diphosphate; GGPPSase, geranylgeranyl diphosphate synthase; HexPPSase, hexaprenyl diphosphate synthase; IPP, isopentenyl diphosphate.

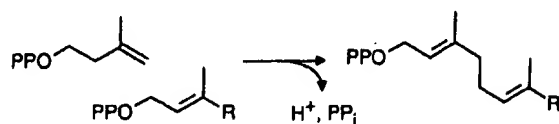


Fig. 1. Synthesis of linear isoprenoid diphosphates from IPP and DMAPP by isoprenyl diphosphate synthases.

bination of function and organisms. We also propose a common α -helical secondary structure for the 13 enzymes.

Results

Pairwise comparisons

Amino acid sequences for 6 FPPSases, 6 GGPPSases, and a HexPPSase were compared pairwise by the Needleman and Wunsch method using the TREE program of Feng and Doolittle (1987, 1990), and the results are shown in Table 1. There was 45–84% amino acid identity among the eukaryotic FPPSases from humans (Sheares et al., 1989), rats (Clarke et al., 1987), chickens (Kroon, unpubl. results), and yeast (Anderson et al., 1989). Substantially lower identities of 10–25% were seen when the eukaryotic FPPSases were compared as a group with the other 9 prenyltransferases, including the eubacterial FPPSases from *Escherichia coli* (Fujisaki et al., 1990) and *Bacillus stearothermophilus* (Koyama et al., 1993). However, the 2 eubacterial FPPSases showed substantial identities of 27–44% with the chromoplast (chloroplast-related) GGPPSase from *Capsicum annuum* (Kuntz et al., 1992), the bifunctional archaeobacterial FPP/GGPPSase from *Methanobacterium thermoautotrophicum* (Chen & Poulter, unpubl. results), and the eubacterial GGPPSases from *Erwinia herbicola* (Armstrong et al., 1990; Math et al., 1992), *Erwinia uredovora* (Misawa et al., 1990), and *Rhodobacter capsulatus* (Armstrong et al., 1989). The fungal

GGPPSase from *Neurospora crassa* (Carattoli et al., 1991) and yeast HexPPSase (Ashby & Edwards, 1990) had lower sequence identities of 10–25% with the other prenyltransferases.

Multiple sequence alignments, conserved sequences, and a phylogenetic tree for isoprenyl diphosphate synthase

The amino acid sequences for 13 isoprenyl diphosphates were aligned according to the procedures of Feng and Doolittle (1990), as shown in Figure 2. Five regions, designated I–V, were found where highly conserved residues appeared in at least 12 of the 13 sequences. Similar regions were identified by Koyama et al. (1993) for a more limited set of isoprenyl diphosphate synthases. Regions II and V are rich in negatively charged aspartates and positively charged arginines or lysines. These sequences correspond to those originally labeled as domains I and II, respectively, by Ashby et al. (1990), who proposed they were diphosphate binding motifs.

The relatively high pairwise percentage identities among selected pairs of prenyltransferases listed in Table 1 are consistent with all 13 isoprenyl diphosphate synthases having diverged from a common ancestral enzyme. This hypothesis is further supported by the existence of 5 highly conserved regions, 2 of which are of considerable length. However, it was apparent from inspection of the alignments that there is a high degree of divergence between the eukaryotic and eubacterial FPPSases and between the FPPSase and HexPPSase in yeast.

A phylogenetic tree (see Fig. 3) was constructed for the 13 isoprenyl diphosphate synthases using the progressive multiple alignments (Feng & Doolittle, 1987, 1990) shown in Figure 2. Three major groupings were obtained. The most primitive branch was a functional segregation of the higher chain length yeast HexPPSase from the shorter chain farnesyl and geranylgeranyl synthases. The shorter chain length enzymes further segregated into bacterial (eubacteria and archaeobacteria) and eukaryotic clusters. The single exception to this pattern was the

Table 1. Pairwise percent identity of isoprenyl diphosphate synthases^a

Protein ^b	FPP_HUM	FPP_RAT	FPP_CHI	FPP_YSC	FPP_ECO	RPP_BST	GGPP_CAN	GGPP_MTH	GGPP_EHE	GGPP_EUR	GGPP_RCA	GGPP_NCR	HPP_YSC
FPP_HUM	—	84.1	68.0	45.0	23.1	22.3	17.6	21.2	18.9	18.4	18.4	12.6	16.1
FPP_RAT		—	66.0	46.0	24.8	22.7	17.6	23.0	19.5	18.4	19.2	13.1	17.0
FPP_CHI			—	46.5	22.0	24.9	18.7	23.5	18.4	18.6	19.8	12.9	16.3
FPP_YSC				—	21.2	22.6	20.7	23.9	19.4	15.7	20.2	10.3	17.0
FPP_ECO					—	42.9	34.7	30.8	31.7	31.9	29.4	15.8	23.6
FPP_BST						—	39.6	33.4	33.6	33.6	31.3	12.9	25.6
GGPP_CAN							—	28.7	27.1	27.1	25.8	12.5	20.4
GGPP_MTH								—	25.7	26.6	29.9	16.1	22.7
GGPP_EHE									—	51.8	26.4	11.8	21.8
GGPP_EUR										—	25.6	11.3	22.5
GGPP_RCA											—	13.9	20.6
GGPP_NCR												—	10.7
HPP_YSC													—

^a Pairwise sequence comparisons were done by TREE of Feng and Doolittle (1987). The percent identity was based on the aligned regions.

^b FPP_HUM, *Homo sapiens* FPPSase; FPP_RAT, *Rattus rattus* FPPSase; FPP_CHI, *Gallus gallus* FPPSase; FPP_YSC, yeast *Saccharomyces cerevisiae* FPPSase; FPP_ECO, *Escherichia coli* FPPSase; FPP_BST, *Bacillus stearothermophilus* FPPSase; GGPP_CAN, *Capsicum annuum* GGPPSase; GGPP_MTH, *Methanobacterium thermoautotrophicum* GGPPSase; GGPP_EHE, *Erwinia herbicola* GGPPSase; GGPP_EUR, *Erwinia uredovora* GGPPSase; GGPP_RCA, *Rhodobacter capsulatus* GGPPSase; GGPP_NCR, *Neurospora crassa* GGPPSase; HPP_YSC yeast *S. cerevisiae* HexPPSase.

FPP_HUM	MNGDQNSDVYAQEKQDFVQHFSQIVRVLTEDMGHPEIGDAIARLKEVLEYNA	IGGKYNRGLTVVAFRELVEPRKQD	ADSLQRAMTVGWCVEL	94		
FPP_RAT	MNGDQKLDVHNQEKQNFQHFQIVKVLTEDELGHPEKGDATRIKEVLEYNT	VGGKYNRGLTVVQTFQELVEPRKQD	AESLQRALTQWCVEL	94		
FPP_CHI	M (17) LSPVVEREREEFVGFFPQIVRDLTEDEGIGHPEVGDAVARLKEVLEYNA	PGGKYNRGLTVVAAVRELSPGQKD	AESLRCAVAVGWCIEL	108		
FPP_YSC	MASEKEIRREERFLNVFPKLVEELNASLLAYGNPKACDNYAHSNLNYNT	PGGKYNRGLSVVDYTAIILSNKTVQEL	GQEEYEKVAIGWCIEL	91		
FPP_ECO	MDFPQQLQACVKQANQALSRIAPLPFQNTPVVETMAYGALLGGKRLRFLVYATGHMFG		VSTNTLDAPAAVVEC	75		
FPP_BST	MAQLSVEQFLNEQKQAVETALSRYIERLEGPALKKAMAYSLEAGGKRIPLLLSTVRAIG		KDPVAVGLPVACAIEM	77		
GGPP_CAN	M (63) ERIEAAQTEEPFNFKIYVTEKASVNAKDEAIIVKEPFIHEAMRYSLEAGGKRVFMLCLAAECVLG		GNQENANAAACAVEM	148		
GGPP_MTH	MTEVLDILRKYSVADKRIMECISDITPTTLKASEHLITAGGKRIPLSALLSCEAVG		GNPDAAGVAAIEL	74		
GGPP_EHE	MVSGSKAGVSPHREIEVMROSIDHLAGLLPETSQDITVSLAMREGVMAPGKRIPLMLLAARDI.RY		QGSMPITLLDLACAVEL	84		
GGPP_EUR	MTVCAKKHVLHTRDAEAQLLADIRRLDQLLPVEGERDVGGAAMREGALAPGKRIPLMLLLTARDLGC		AVSHOGLLDLACAVEN	85		
GGPP_RCA	MSLDKRIESALVKALSPEALGESPPLLAALPVGVEGGARIRPTILVSVALACG		DDCPAVTDAAVALEL	71		
GGPP_NCR	M (87) FSPYTHAPQPPPPPPNPPORFATEDFFSPSRRTTSEEKEKVLTPGYDYLNGHPGKDRIQSNVKAFAWLD		VPSESLVETIKVISM	172		
HPP_YSC	M (29) AASKLVTPKILMNNPISLVSKENMTLAKNIVALIGSGHPVLNKVTSYFTEGKKVRPLLVLLSRALS	(70)	GILPKQRRALAEIVEM	184		
Consensus		<u>GK...R</u>	<u>E</u>			
I						
FPP_HUM	LQAFFLVADDI	MDSSLTRRQQTCTWYQKPGVGLDAINDANLLACIYRLKL	YCREQPYLNLIELFLQSSYQTEI	QGTLDLLTAPQGNVDLVR	187	
FPP_RAT	LQAFFLVDDI	MDSSYTRRQQTCTWYQKPGIGLDAINDALLAAIYRLKL	YCREQPYLNLIELFLQSSYQTEI	QGTLDLITAPQGNVDLGR	187	
FPP_CHI	FQAFFLVADDI	MDQSLTRRQQLCTWYKQEGVGLDAINDSDLESSVYRLKL	YCRQRPYVHLELFLQATAYQTEL	QGMLDLITAPVSKVDLSH	201	
FPP_YSC	LQAFFLVADDI	MDKSI TRRQQTCTWYKQVPEVGEIANDAFMLAAIYRLKL	HFNEKYIYDITELFHEVTQTEL	QGLMDLTAPVSKVDLSH	184	
FPP_ECO	IBAYSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		MCGGAALDLOAGKGVFLDA	173		
FPP_BST	IBYSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		MVAGQAADMEGEGKT	LTLE	175	
GGPP_CAN	INTMSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		LVAGQVADIKCTGNASVLEI	246		
GGPP_MTH	INTMSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		ICGQALDMEGFEERLDVTE	165		
GGPP_EHE	INTMSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		LVLGQFRDLNDAALDRTPDA	180		
GGPP_EUR	INTMSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		LVQGOFFDLSEKQVPSA	181		
GGPP_RCA	INTMSLIHDDLPAMDNDLRRGKPTNRKVFGEAMAILAGDGLTYAFQILITEIDDERIPPSVRLRLTERLAKAAGPEG		LHRGOGHDLFWRDTLCTPTD	259		
GGPP_NCR	INTMSLIHDDVP	EDNSVLRRGFPVARSIFGIPQITNTSNVYVYFALQVQLKLNKPAYSIPELLN				
HPP_YSC	INTMSLIHDDV	IDHSOTRRGRPSGNAFTNKNMAVLGAGDFLLGRATVSIIRLHNPEVVELMHSNIAINLVE	(33)	KEHDFRVPSSROGLQLSHDQITE	310	
Consensus	<u>....L...D...D...RRG</u>		<u>GQ...D</u>			
II						
FPP_HUM	FTEKRYKSIYKTAIFYSYFLPIAAAHYMAIGDGEKEHANAKKILLEMGEFFQIQQDDYLDLFGDPSVTGK	IGTDIQDNKCSMLVQCLQRATPEQYQIL	286			
FPP_RAT	FTEKRYKSIYKTAIFYSYFLPIAAAHYMAIGDGEKEHANAKKILLEMGEFFQIQQDDYLDLFGDPSVTGK	IGTDIQDNKCSMLVQCLQRATPEQYQIL	286			
FPP_CHI	FSEERYKAIYKTAIFYSYFLPVAAAHYMGIDSKKEHENAKKILLEMGEFFQIQQDDYLDLFGDPSVTGK	IGTDIQDNKCSMLVQCLQRATPEQYQIL	300			
FPP_YSC	FSLKXHSFIVTFKTAIFYSYFLPVAAAHYMAIGDGEKEHANAKKILLEMGEFFQIQQDDYLDLFGDPSVTGK	IGTDIQDNKCSMLVQCLQRATPEQYQIL	283			
FPP_ECO	LE RIRHKTGAL	IRAAVRLGALSAGDKGRALPLVDKYAESGLAFQVDDILDVVGDTATLGRQGAQQGLKSTYPAALLGLEQARKKARDI	267			
FPP_BST	LE YIHRKTKM	LQISVHAGALIGCADARQT	RELDEFRAHGLAFQIRDDILDIEGAEKIKGKPVGSDQSNKATYPALLSLAGAKELAFHI	268		
GGPP_CAN	LE FIVHKTAL	LESSVVLGAILGGG	THVEVLEKRLRFACIGLGLFQVDDILDVVGDTATLGRQGAQQGLKSTYPAALLGLEQARKKARDI	267		
GGPP_MTH	YME MIYK KTAAL	IAAATKAGAIMGASER	EVEALEDYKGFGLAFQIRDDILDVVGDTATLGRQGAQQGLKSTYPAALLGLEQARKKARDI	267		
GGPP_EHE	IL STNHLKTGIL	FSAMLOIVAIASASSPSTR	ETLHAFALDFGQAFQLDDLDLDDHPT	GKDRNKD	AGKSTLVNRLGADAARQKREHI	268
GGPP_EUR	IL MTNHFKTSL	FCASMQMASIVANASSEAR	DCLHRFSLDLQGFQAFQLDDLDLDDHPT	GKDSNDQ	AGKSTLVNRLGADAARQKREHI	268
GGPP_RCA	LA AYHQAKTAL	PIAATQMGAIAGYAEAPWF	LGRIGSAFQIADDLKALMSAEAMGKPAQODIANERPNVAKTMGIEGARKHLQDVL	252		
GGPP_NCR	DYL EMVSNKTGGL	PRIGIKMQAESRSPVDCVP	LVNITGLIFQIADDDYHNLNREYATANKMCEDLTGKFSFPVHSIRSNPSNMQLN	249		
HPP_YSC	TAFEYIHKTYLKTAL	ISKSCRCAAILSGASPAVI	DECYDFGRNLGICFQLVDDMLDFTVSGKDLGKPSGADLKLGTATAPVLFAMKEDPSLGLTIS	408		
Consensus	<u>KT</u>		<u>G...FQ...DD...D...GK...D...K</u>			
III						
FPP_HUM	KENYQKEAEKVARVKALYEELDPAVFLQYEEDSYSHIALIEQYAAP	LPPAVFLGLARKIYKRRK	353			
FPP_RAT	EENYQKQDPEKVARVKALYEELDPAVFLQYEEDSYSHIALIEQYAAP	LPPSIFLELANKIYKRRK	353			
FPP_CHI	EDNYGRKEPEKVARVKALYEELDPAVFLQYEEDSYSHIALIEQYAAP	LPKEIFLGLAQKIYKRRK	367			
FPP_YSC	DENYGKDSVAEAKCKKIFNDLKTIEQLYHEYESIAKDLKAKISQVDESARGFADVLTAFLNKVYKRSK		352			
FPP_ECO	DDAROSLQQLAEQSLDTSALALADYIQRNK		299			
FPP_BST	EAAQRHLRNADVDGAALAYICELVAARDH		297			
GGPP_CAN	REAKQQLGFDPSRKAAPLIALADYIARDN		369			
GGPP_MTH	ISILSGDEGSVAEAEIFERY	GATQYAEVALDYVRMAKERLEILEDSDARDA	LMRIADPVLREH	325		
GGPP_EHE	DSADKHLTFACPGGAIQFMHLMFGHHLADNSPVMKIA		307			
GGPP_EUR	QLASEHLISACQGHATQHTQAFDPKKLAAYS		302			
GGPP_RCA	AGAIAISIPSCPGKALQAHQVLYAHKIMDIPASAEERG		289			
GGPP_NCR	ILKQKTGDZEVKRYAVAYMESTGSEYTRKVIKLVDRARQHTEDIDGGRKSGGIHILDRIMHQQENVAQKNGKKE		428			
HPP_YSC	RNFSERGDVEKTIQSVRLHNGIAKTKILAEZYDKALQNLRLSPESDARSALZFLTNSILTRRK		473			

Fig. 2. Multiple sequence alignment for the 13 isoprenyl diphosphate synthases listed in Table 1. Long N-terminal sequences and insertions in HPP_YSC are omitted, but the numbers of amino acids are shown in parentheses. Consensus sequences shown below the 5 highly conserved sequence domains, I-V, are double underlined. A region clearly corresponding to domain III was not seen in HPP_YSC. Residues conserved differently in eukaryotic FPP synthases are in bold. The peptide in chicken FPP synthase that was labeled during photoaffinity experiments is underlined.

inclusion of the chromoplastic GGPPSase from *C. annuum* (green peppers) in a cluster of eubacterial farnesyl and geranylgeranyl diphosphate synthases. These results indicate that the chain length selectivity of the short-chain prenyltransferases cannot be readily deduced from sequence comparisons and that assignments of function should be made biochemically.

Prediction of secondary structure

Because the pairwise alignments indicate that the isoprenyl diphosphate synthases have diverged from a common ancestor, it is reasonable to assume that the gross topological features of the ancestor were conserved during evolution. We initially compared the amino acid compositions of 11 prenyltransferases, all of the FPPSases and GGPPSases except the highly diverged *N. crassa* enzyme, using Chou's approach (Chou, 1989) for predict-

ing structural classes of proteins from their amino acid contents. The results are shown in Table 2. As judged by comparing the average amino acid composition between the isoprenyl diphosphate synthases with those of representative all- α , all- β , $\alpha + \beta$, and α/β proteins, the prenyltransferases most closely resemble typical all- α or α/β structures, suggesting an all α -helix protein or a protein dominated by α -helices.

A consensus secondary structure for the isoprenyl diphosphate synthases was predicted from a combination of the multiple sequence alignments, probabilities for formation of loop, α -helix, and β -sheet regions (Fig. 4), and an average hydropathy plot (Fig. 5). Predictions by Garnier-Osguthorpe-Robson (GOR) or Chou-Fasman (CF) methods were in good agreement and predicted 8 α -helices and 4 short β -sheets. The location of the α -helices and β -sheets generally correlated well with the average hydropathy plot for the 11 amino acid sequences. Loops 1, 2,

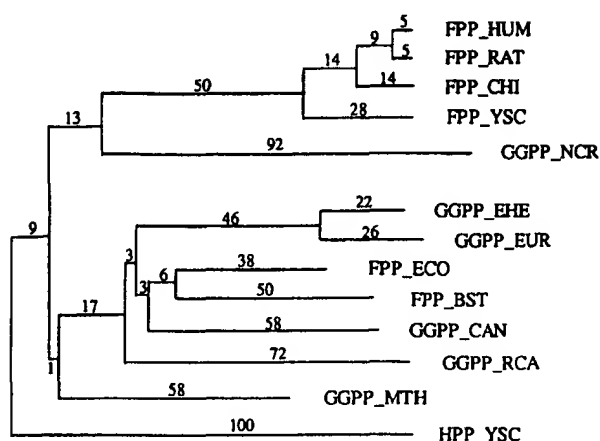


Fig. 3. A phylogenetic tree for isoprenyl diphosphate synthases constructed from progressive alignments using the TREE program and refined as described by Feng and Doolittle (1990).

3, 4, and 7 were consistently assigned by computer predicted turns, gaps in the alignments, and hydrophilic peaks in the hydropathy plot. A short β -sheet was predicted within loop 3 and within 7 by GOR and CF algorithms. However, the hydropathy plots placed these "sheet" sequences in hydrophilic regions and cast doubt on their existence. The assignments for loops 5 and 6 were based on large gaps that occurred in these regions and large negative hydropathy indices. The assignment for loop 8 was based on large negative hydropathy indices in that region. There were also gaps in the alignment between $\alpha 4$ and $\alpha 5$; however, this region was not hydrophilic, and a turn motif was not predicted by GOR or CF. Perhaps these 2 helices are joined by a spacer of variable length or are fused into a single α -helix. In addition, no turns or loops were predicted between $\beta 2$ and $\alpha 6$ or $\alpha 7$ and $\beta 3$. Because the average amino acid composition predicted a structure primarily composed of α -helices, the short $\beta 2$ and $\beta 3$ regions may be helical extensions of $\alpha 6$ and $\alpha 7$, respectively, rather than β -sheets.

A predicted average secondary structure for the isoprenyl diphosphate synthases is presented in Figure 6. The high α -helix content in the structure is consistent with the statistical prediction based on amino acid composition. The secondary structural elements were arranged to place the 5 regions containing highly conserved sequences together on the same face of the structure. Although the 3-dimensional fold is not known for any prenyltransferase, one might imagine an antiparallel orientation of $\alpha 2$, $\alpha 3$, and $\alpha 4/\alpha 5$ that allows loops 3, 5, and 7 to be brought together. Additional support for this folding pattern is discussed in the next section. Because the consensus structure was constructed from homologous core sequences, individual enzymes may contain some additional elements of secondary structure that lay outside of the predicted consensus regions. Likewise, the lengths of some secondary structural elements, loops, and spacers undoubtedly vary from protein to protein.

A model for substrate binding

The predicted consensus structure, along with other information about catalytic site residues, can serve as a guide for locat-

Table 2. Comparison of average amino acid compositions of isoprenyl diphosphate synthases and the 4 protein classes^a

Amino acid	Synthases ^b	All- α	All- β	$\alpha + \beta$	α/β
Ala	11.2	11.6	7.3	9.3	8.3
Arg	5.1	2.2	2.4	4.1	3.4
Asn	2.5	4.0	5.0	6.4	4.2
Asp	7.0	6.7	4.4	5.9	5.6
Cys	1.6	0.9	2.7	3.9	1.5
Gln	4.6	2.7	4.4	3.9	2.6
Glu	7.3	5.5	3.1	4.6	5.9
Gly	6.7	8.1	10.7	9.1	8.7
His	2.3	4.5	1.8	1.7	2.5
Ile	5.5	3.7	4.3	4.9	5.5
Leu	11.7	9.0	6.4	5.8	7.8
Lys	6.1	10.2	4.1	5.9	7.4
Met	2.6	2.0	0.6	1.3	2.1
Phe	3.3	5.0	3.1	2.8	3.6
Pro	3.5	3.4	4.6	3.8	4.3
Ser	3.1	5.0	12.3	6.7	7.5
Thr	3.9	4.9	9.1	6.2	5.5
Trp	0.6	1.3	1.6	1.6	1.7
Tyr	3.3	2.6	4.0	5.7	3.0
Val	6.0	6.8	8.2	6.5	8.7
Difference index ^c		27.4	49.8	33.2	28.3

^a Values for protein class all- α , all- β , $\alpha + \beta$, and α/β are taken from Chou (1989).

^b Average amino acid compositions of all FPP synthases and GGPP synthases (not including GGPP_NCR).

^c Difference index = $\sum |C_{Ai} - C_{Bi}|$ as described in Methods.

ing putative binding sites for the substrates. The 5 highly conserved regions identified in Figure 2 are located in the secondary structure as follows (see Figs. 4, 6): Region I, from loop 1 to the N-terminal part of $\beta 1$; Region II, from the C-terminal half of $\alpha 2$ to the N-terminal half of loop 3; Region III, $\alpha 5$; Region IV, the C-terminus of loop 5; and Region V, from the C-terminus of $\alpha 7$ through loop 7. Photoaffinity experiments with an azido analog of IPP (Brems et al., 1981) labeled several amino acids from positions 157 to 188 in L5 of avian FPPSase, suggesting that this part of the enzyme interacts with the hydrophobic isopentenyl moiety in IPP. Recently, Blanchard and Karst (1993) discovered that a mutation at K197 near the C-terminus of L5 in yeast FPPSase both reduced the activity and altered chain length selectivity of the enzyme. K197 is located just beyond the region labeled in the photoaffinity studies with avian protein. These results suggest that much of L5 forms an integral part of the IPP pocket with the C-terminal end of the loop extending to the binding site for the allylic substrate.

The highly conserved DDXX(XX)D motifs in L3 and L7, as well as the arginine doublet in L3, are likely candidates for diphosphate binding sites. These predictions (Ashby et al., 1990) are consistent with site-directed mutagenesis experiments (Joly & Edwards, 1993; Song & Poulter, 1994) that established that all of these residues except the last aspartate in L7 were essential for catalysis. Which substrate binds to which aspartate-rich region is not known. Ashby et al. (1990) suggested that the DDXXD motif in L7 is the allylic binding site on the basis of sequencing comparisons with prenyltransferases that utilize non-

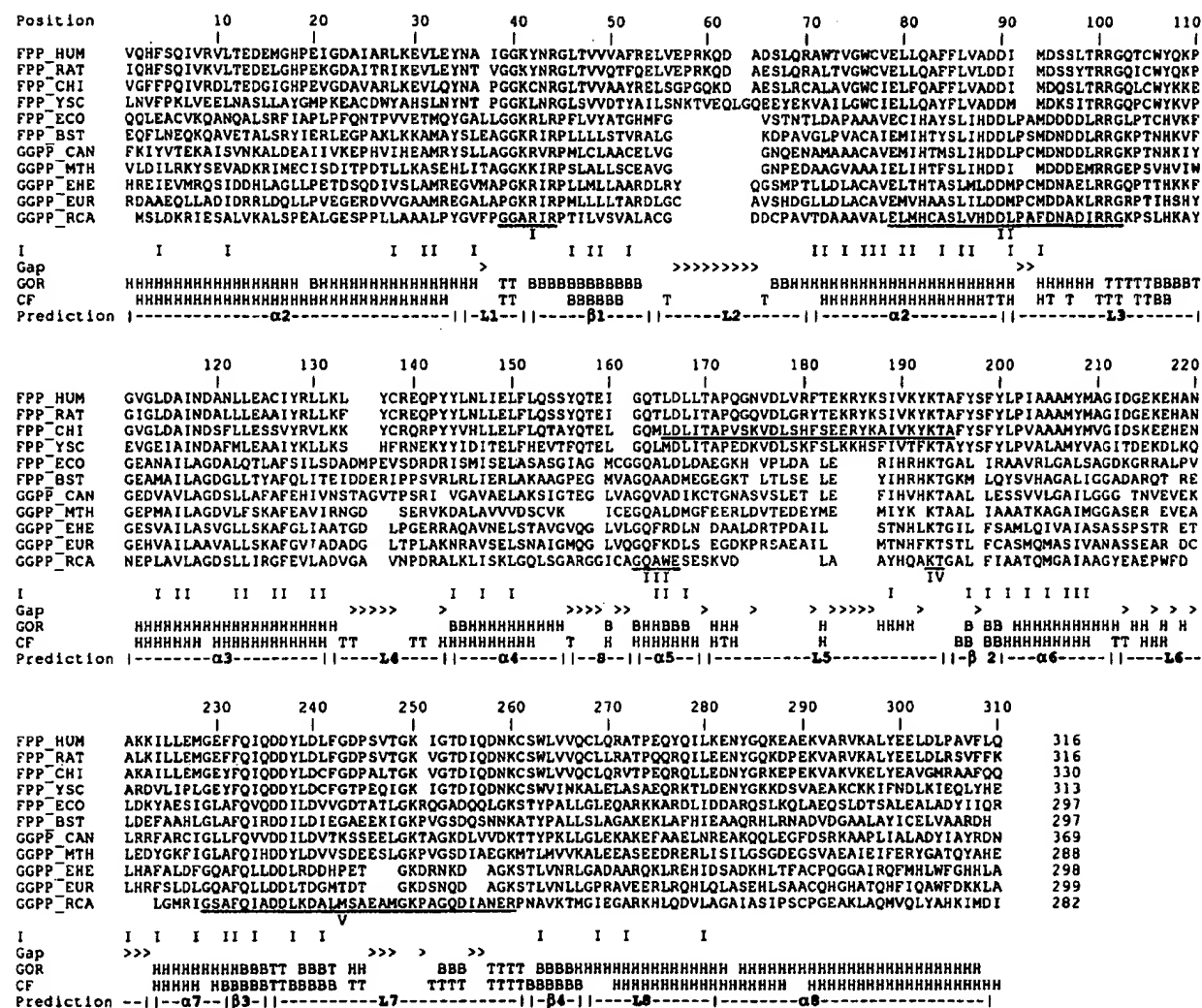


Fig. 4. A predicted consensus secondary structure based on multiple sequence alignments. The alignment of FPP synthases and GGPP synthases (not including GGPP_NCR) is the same as in Figure 2 except the N- and C-termini are omitted. The numbers on top of sequences show the alignment positions, whereas the numbers on the right are the actual amino acid numbers. The 5 conserved domains are double underlined, and the affinity labeled peptide in chicken FPP synthase is underlined. I, hydrophobic residues; CF, consensus secondary structure prediction by Chou and Fasman; GOR, consensus secondary structure prediction by GOR. The inserted gaps in the alignment are also marked by >. H and α , α -helix; B and β , β -sheet; L, loop; S, spacer; T, turn.

isoprenoid acceptors instead of IPP; however, except for the 3 aspartates, the overall sequence homologies in this region were low. A helical wheel projection of $\alpha 2$ and $\alpha 3$ indicates that a substantial portion of the total exposed surface area of these helices is hydrophobic. An alternative model for substrate binding has the diphosphate moiety in the allylic substrate interacting with the DDXX(XX)D motif in L3 instead of L7, with $\alpha 2$ and $\alpha 3$ facilitating binding of the hydrophobic isoprenoid tail through hydrophobic interactions. In this scenario, the diphosphate residue in IPP binds to the DDXXD region in L7 with the hydrophobic isopentenyl moiety in the region of the active site bounded by most of L5, which was labeled with the IPP photoaffinity analog.

All of the isoprenyl diphosphate synthases except the GGPPSases from *Erwinia* and *Rhodobacter* have charged side

chains in 2 of the final 3 C-terminal residues. Amino acids containing positively charged side chains appear in the first and third positions in most of the enzymes. Site-directed mutagenesis of R350 in *Saccharomyces cerevisiae* FPPSase had little effect on the catalytic constants for the enzyme (Song & Poulter, 1994). However, fusion of a negatively charged EEF α -tubulin C-terminal epitope to the wild-type sequence reduced V_{max} 12-fold and was accompanied by a 14-fold increase in K_M for IPP. Laskovics and Poulter (1981) measured the individual kinetic constants for avian FPPSase and found that the rates of addition of substrates were substantially below the diffusion-controlled limits. These results are consistent with a conformational change in FPPSase upon binding of substrates. Thus, the C-terminus of the enzyme may form a flexible flap that helps seal the active site during catalysis.

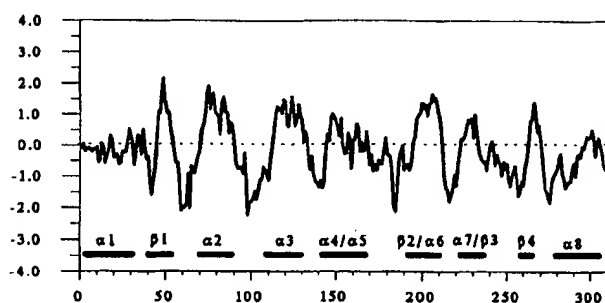


Fig. 5. An average hydropathy plot for isoprenyl diphosphate synthases. Hydropathy indices of FPP synthases and GGPP synthases (not including GGPP_NCR) were averaged at homologous positions according to the alignment shown in Figure 4. The average index is plotted along the alignment positions. Predicted α -helix and β -sheet structures are shown below the plot.

Discussion

Isoprenyl diphosphate synthases catalyze the basic chain elongation steps in the isoprenoid biosynthetic pathway. These reactions are ubiquitous in nature. Organisms contain 2 classes of isoprenyl diphosphate synthases, one for synthesis of short-chain C_{10} – C_{20} molecules and another for longer chain isoprenoids. The short-chain enzymes are further subdivided into specific enzymes for synthesis of geranyl, farnesyl, and geranylgeranyl diphosphate. The long-chain prenyltransferases are also subdivided by chain length selectivity and, in addition, specifically form either *cis* or *trans* double bonds in the newly added isoprene units. Amino acid sequences are now available for several short-chain FPPSases and GGPPSases and for 1 all-*trans* long-chain synthase. Comparisons of the primary sequences for 13 isoprenyl diphosphate synthases shown in Figure 2 revealed 5 regions containing 2–10 highly conserved amino acids. Analysis of multiply aligned sequences for the 11 FPPSases and GGPPases shown in Figure 4, in conjunction with predictions of secondary structures, indicated that these enzymes are all α -helix proteins or α/β structures dominated by α -helices.

Simple inspection of the aligned sequences suggested that individual members of the family diverged from a common an-

cestral isoprenyl diphosphate synthase during evolution (James et al., 1978; Bajaj & Blundell, 1984; Chothia & Lesk, 1986). A more quantitative analysis using the methods of Feng and Doolittle (1990) supported this hypothesis and provided significant insights into the pathway by which they evolved. The earliest branch was a functional segregation that separated the long-chain from the short-chain synthases, as illustrated by the large divergence between the FPPSase and the HexPPSase in yeast.

The second major branching evident in the phylogenetic tree presented in Figure 3 segregates the short-chain length prenyltransferases into 2 clusters regardless of chain length, one for eubacterial and archaeobacterial proteins, and another for eukaryotic enzymes. Many organisms have distinct enzymes for synthesis of C_{10} – C_{20} isoprenyl diphosphates when these compounds serve as substrates for other enzymes. Thus, one might have anticipated a primary clustering for the short-chain enzymes according to chain length rather than kingdom. However, *M. thermoautotrophicum*, a methanogenic archaeobacterium, has a single bifunctional short-chain prenyltransferase that provides both the C_{15} precursor for synthesis of squalene and the C_{20} precursor for synthesis of the distinctive isoprenoid glyceryl ether core membrane lipids found in members of the archae kingdom (Chen & Poulter, 1993). Thus, the archaeobacterial enzyme may represent a primitive scenario where a single enzyme was responsible for short-chain synthesis. In this case, the fine tuning of chain length control would have evolved independently after eukaryotes and eubacteria diverged. Additional examples of eukaryotic GGPP synthases should help clarify this point. The single exception to the clustering pattern for the short-chain synthases is the chromoplastic GGPPSase from peppers, where the gene for the enzyme may have been captured from an ancient bacterial symbiote.

It is unclear what mechanism regulates how many molecules of IPP are added to the growing isoprenoid chain by a prenyltransferase, and there appear to be no clues from the amino acid sequences shown in Figure 2. This question will not be resolved until more sequence information is available or X-ray structures are obtained for prenyltransferases with different chain-length selectivities.

Although the correlations we discovered provide important clues about the evolution of isoprenyl diphosphate synthases,

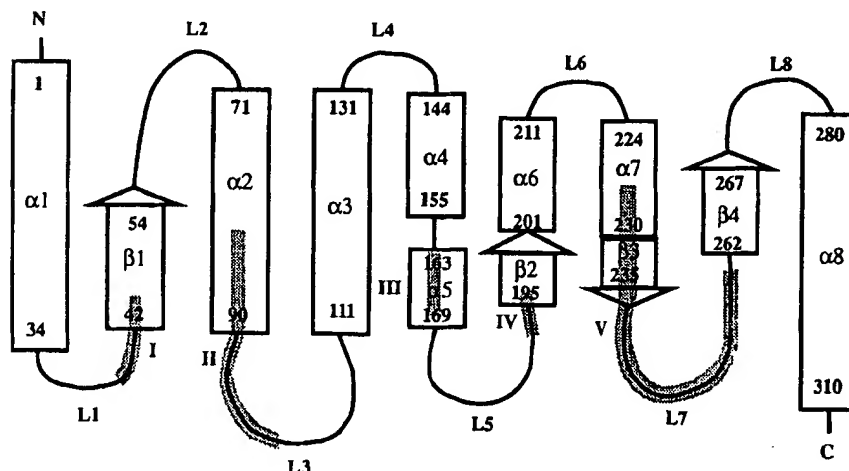


Fig. 6. The predicted average secondary structure of FPP synthases and GGPP synthases. α -Helices ($\alpha 1$ – $\alpha 8$) and β -sheets ($\beta 1$ – $\beta 4$) are drawn in rectangles and arrows, respectively. Loops (L1–L8) are shown as curved lines. Each secondary structural unit is numbered by its position in the alignment shown in Figure 4 (see text for alternative views on $\alpha 4$ – $\alpha 7$ and $\beta 2$ – $\beta 3$). The 5 conserved domains (shaded) are labeled I–V. The drawing is not to scale.

the phylogenetic tree is not complete. There are no sequences yet reported for a long-chain *cis* double-bond synthase, and more examples of eukaryotic GGPP synthases are needed to confirm the groupings we propose. With the high level of activity in this area at present, these gaps should be filled in the near future.

Methods

The protein sequences of all isoprenyl diphosphate synthases except FPP_BST (Koyama et al., 1993) and FPP_CHI (Kroon et al., unpubl.) were retrieved from the Swiss-Prot data bank using GCG programs (University of Wisconsin Genetics Computer Group). The TREE program (Feng & Doolittle, 1990) was used for pairwise comparisons, to perform multiple sequence alignments, and to construct a phylogenetic tree. Refinements in the tree were made according to the protocols described by Feng and Doolittle (1990).

Average amino acid compositions of 6 FPP synthases and 6 GGPP synthases (all except GGPP_NCR) were calculated using a spreadsheet. The difference index between the average amino acid compositions of the isoprenyl diphosphate synthases and those of all- α , all- β , $\alpha + \beta$, and α/β proteins were the sum of composition differences for each amino acid, $\sum |C_{Ai} - C_{Bi}|$ (Chou, 1989; Doolittle, 1992). The smaller the difference index, the closer the comparison.

To predict an average secondary structure for FPPSases and GGPPSases, a secondary structure was calculated for each protein by the GOR procedure (Garner et al., 1978) using GARNER in PCGENE and the CF method (Chou & Fasman, 1974) using PEPTIDESTRUCTURE in GCG. The secondary structures were then arranged according to a multiple alignment truncated at both N- and C-termini. Consensus assignments of α -helix, β -sheet, or turn structures to each alignment position were determined when the assigned structural feature appeared in more than half of the aligned sequences, except at positions where gaps were inserted.

The multiple sequence alignment itself may provide information about turns and surface loops. Regions where gaps were inserted were normally considered as surface loops to accommodate insertion or deletion of a few amino acids. Additional information about structure came from hydropathy plots where regions of high hydrophobicity usually correlate with a buried β -sheet or a hydrophobic α -helix, whereas regions of high hydrophilicity usually correlate with a surface loop or a turn. An amphipathic α -helix normally occurs where there is no high or low peak in hydropathy plot. Hydropathy indices were calculated for each prenyltransferase by the method of Kyte and Doolittle (1982) using PEPTIDESTRUCTURE. These values were averaged to calculate a hydropathy index at the corresponding positions in the multiple alignment. The average indices were plotted along the alignment positions. The consensus secondary structure was predicted by combining GOR and CF structures, consideration of gaps in the alignment, and comparisons with the average hydropathy plot. The hydrophobic nature of side chains at positions containing L, I, V, M, F, W, Y, A, or C in at least 9 of 11 sequences was also indicated as I (for interior) to help visualize the hydrophobicity of secondary structure units. Helical wheel projections were constructed by HELWHEEL in PCGENE to facilitate analysis of the surface of the α -helices in the consensus structure (Shiffer & Edmundson, 1967).

Acknowledgments

We thank R.F. Doolittle for providing copies of his program and for helpful discussions. This work was supported by NIH grant GM 21328.

References

- Allen CM. 1985. Purification and characterization of undecaprenyl pyrophosphate synthetase. *Methods Enzymol* 110:281-299.
- Anderson MS, Yarger JG, Burck CL, Poulter CD. 1989. Farnesyl diphosphate synthetase. Molecular cloning, sequence, and expression of an essential gene from *Saccharomyces cerevisiae*. *J Biol Chem* 264:19176-19184.
- Armstrong GA, Alberti M, Hearst JE. 1990. Conserved enzymes mediate the early reactions of carotenoid biosynthesis in nonphotosynthetic and photosynthetic prokaryotes. *Proc Natl Acad Sci USA* 87:9975-9979.
- Armstrong GA, Alberti M, Leach F, Hearst JE. 1989. Nucleotide sequence, organization, and nature of the protein products of the carotenoid biosynthesis gene cluster of *Rhodobacter capsulatus*. *Mol Gen Genet* 216:254-268.
- Ashby MN, Edwards PA. 1990. Elucidation of the deficiency in two yeast coenzyme Q mutants. *J Biol Chem* 265:13157-13164.
- Ashby MN, Spear DH, Edwards PA. 1990. Prenyltransferases from yeast to man. In: Attie AD, ed. *Molecular biology of atherosclerosis*. Amsterdam: Elsevier Science Publishers. pp 27-34.
- Bajaj M, Blundell T. 1984. Evolution and the tertiary structure of proteins. *Annu Rev Biophys Bioeng* 13:453-492.
- Blanchard L, Karst F. 1993. Characterization of a lysine-to-glutamic acid mutation in a conservative sequence of farnesyl diphosphate synthase from *Saccharomyces cerevisiae*. *Gene* 125:185-189.
- Brems DN, Bruenger E, Rilling HC. 1981. Isolation and characterization of a photoaffinity-labeled peptide from the catalytic site of prenyltransferase. *Biochemistry* 20:3711-3718.
- Brems DN, Rilling HC. 1979. Photoaffinity labeling of the catalytic site of prenyltransferase. *Biochemistry* 18:860-864.
- Cane DE. 1981. Biosynthesis of sesquiterpenes. In: Porter JW, Spurgeon SL, eds. *Biosynthesis of isoprenoid compounds, vol 1*. New York: John Wiley & Sons. pp 283-374.
- Carattoli A, Romano N, Ballario P, Morelli G, Macino G. 1991. The *Neurospora crassa* carotenoid biosynthetic gene (Albino 3) reveals highly conserved regions among prenyltransferases. *J Biol Chem* 266:5854-5859.
- Chen A, Poulter CD. 1993. Purification and characterization of farnesyl diphosphate/geranylgeranyl diphosphate synthase. A thermostable bifunctional enzyme from *Methanobacterium thermoautotrophicum*. *J Biol Chem* 268:11002-11007.
- Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823-826.
- Chou PY. 1989. Prediction of protein structural classes from amino acid compositions. In: Fasman GD, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 549-586.
- Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry* 13:222-224.
- Clarke CF, Tanaka RD, Svenson K, Wamsley M, Fogelman AM, Edwards PA. 1987. Molecular cloning and sequence of a cholesterol-repressible enzyme related to prenyltransferase in the isoprene biosynthetic pathway. *Mol Cell Biol* 7:3138-3146.
- Clarke S. 1992. Protein isoprenylation and methylation at carboxyl-terminal cysteine residues. *Annu Rev Biochem* 61:355-386.
- Crawford IP, Niermann T, Kirschner K. 1987. Prediction of secondary structure by evolutionary comparison: Application to the α subunit of tryptophan synthase. *Protein Struct Funct Genet* 2:118-129.
- Croteau R. 1981. Biosynthesis of monoterpenes. In: Porter JW, Spurgeon SL, eds. *Biosynthesis of isoprenoid compounds, vol 1*. New York: John Wiley & Sons. pp 225-282.
- Doolittle RF. 1992. A detailed consideration of a principal domain of vertebrate fibrinogen and its relatives. *Protein Sci* 1:1563-1577.
- Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* 25:351-360.
- Feng DF, Doolittle RF. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol* 183:375-387.
- Fujisaki S, Hara H, Nishimura Y, Horiuchi K, Nishino T. 1990. Cloning and nucleotide sequence of *ispA* gene responsible for farnesyl diphosphate synthase activity in *Escherichia coli*. *J Biochem* 108:995-1000.
- Garner J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 120:97-120.
- James MNG, Delbaere LTJ, Brayer GD. 1978. Amino acid sequence align-

- ment of bacterial and mammalian pancreatic serine proteases based on topological equivalences. *Can J Biochem* 56:396-402.
- Joly A, Edwards PA. 1993. Effect of site-directed mutagenesis of conserved aspartate and arginine residues upon farnesyl diphosphate synthase activity. *J Biol Chem* 268:26983-26989.
- Koyama T, Obata S, Osabe M, Takeshita A, Yokoyama K, Uchida M, Nishino T, Ogura K. 1993. Thermostable farnesyl diphosphate synthase of *Bacillus stearothermophilus*: Molecular cloning, sequence determination, over production, and purification. *J Biochem* 113:355-363.
- Kuntz M, Roemer S, Suire C, Huguency P, Well JH, Schantz R, Camara B. 1992. Identification of a cDNA for the plastid-located geranylgeranyl pyrophosphate synthase from *Capsicum annuum*: Correlative increase in enzyme activity and transcript level during fruit ripening. *Plant J* 2:25-34.
- Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105-132.
- Laskovics FM, Poulter CD. 1981. Prenyltransferase: Determination of the binding mechanism and individual kinetic constants for farnesyl pyrophosphate synthetase by rapid quench and isotope partitioning experiments. *Biochemistry* 20:1893-1901.
- Marrero PF, Poulter CD, Edwards PA. 1992. Effects of site-directed mutagenesis of the highly conserved aspartate residues in domain II of farnesyl diphosphate synthase activity. *J Biol Chem* 267:21873-21878.
- Math SK, Hearst JE, Poulter CD. 1992. The *crtE* gene in *Erwinia herbicola* encodes geranylgeranyl diphosphate synthase. *Proc Natl Acad Sci USA* 89:6761-6764.
- Matsuoka S, Sagami H, Kurisaki A, Ogura K. 1991. Variable product specificity of microsomal dehydrodolichyl diphosphate synthase from rat. *J Biol Chem* 266:3464-3468.
- Misawa N, Nakagawa M, Kobayashi K, Yamano S, Izawa Y, Nakamura K, Harashima K. 1990. Elucidation of the *Erwinia uredovora* carotenoid biosynthetic pathway by functional analysis of gene products expressed in *Escherichia coli*. *J Bacteriol* 172:6704-6712.
- Myers EW, Miller W. 1988. Optimal alignments in linear space. *Comput Appl Biosci* 4:11-17.
- Poulter CD, Rilling HC. 1978. The prenyltransferase reaction. Enzymatic and mechanistic studies of the 1'-4 coupling reaction in the terpene biosynthetic pathway. *Acc Chem Res* 11:307-313.
- Poulter CD, Rilling HC. 1981a. Conversion of farnesyl pyrophosphate to squalene. In: Porter JW, Spurgeon SL, eds. *Biosynthesis of isoprenoid compounds, vol 1*. New York: John Wiley & Sons. pp 225-282.
- Poulter CD, Rilling HC. 1981b. Prenyltransferases and isomerase. In: Porter JW, Spurgeon SL, eds. *Biosynthesis of isoprenoid compounds, vol 1*. New York: John Wiley & Sons. pp 161-224.
- Schiffer M, Edmundson AB. 1967. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys J* 7:121-135.
- Sheares BT, White SS, Molowa DT, Chan K, Ding VDH, Kroon PA, Bostedor RG, Karkas JD. 1989. Cloning, analysis, and bacterial expression of human farnesyl pyrophosphate synthetase and its regulation in Hep G2 cell. *Biochemistry* 28:8129-8135.
- Song LS, Poulter CD. 1994. Yeast farnesyl diphosphate synthase. Site-directed mutagenesis of residues in highly conserved prenyltransferase domain I and domain II. *Proc Natl Acad Sci USA*. Forthcoming.
- Spurgeon SL, Porter JW. 1981. Biosynthesis of carotenoids. In: Porter JW, Spurgeon SL, eds. *Biosynthesis of isoprenoid compounds, vol 2*. New York: John Wiley & Sons. pp 1-122.
- West CA. 1981. Biosynthesis of diterpenes. In: Porter JW, Spurgeon SL, eds. *Biosynthesis of isoprenoid compounds, vol 1*. New York: John Wiley & Sons. pp 375-411.

Exhaustive and Iterative Clustering of the Protein Databank

by K. Kelly

Abstract. The unique high-resolution protein chains in the September 1998 edition of the Brookhaven Protein Databank have been subjected to an exhaustive and iterative sequence- and structure-based clustering procedure to produce a database of alignments for homology identification and modeling. A novel feature of the procedure was the use of multiple-sequence, structure-based alignment to validate hypothesized clusters. The resulting database contains fewer than 800 entries. Homology searches against this database, using rigorous sequence-to-group alignment, show high sensitivity and specificity when compared to existing methodologies. Four models were built based on families in the database and were submitted to the recently completed CASP3 competition.

- [Introduction](#)
- [Methods and Materials](#)
- [Results and Discussion](#)
- [References](#)

Introduction

It is not yet possible to predict the 3D structure of an expressed protein from its amino acid sequence alone. Consequently, inferences about the structure and function of new protein sequences are generally drawn based upon comparisons to sequences for which there already exist experimental models. To date, the most useful principle which is applied to guide such searches is the rule that "similar sequence" implies "similar structure," where sequence similarity is understood to be determined by application of the well-established dynamic programming paradigm (Needleman and Wunsch). Generally speaking, if a protein sequence shares more than 25% pairwise similarity with a known structure, it usually also shares at least the broad outlines of the fold topology of the known structure. However, the Protein Databank contains many pairs of similar structures that are remote homologues with less -- sometimes much less -- than 25% pairwise sequence identity. Standard homology searching tools often have difficulty identifying such remote relationships with any confidence, and even when a relationship is identified, the correct alignment of the new sequence to the homologous structure can be difficult to judge.

Researchers are actively working on strategies for improving the sensitivity and selectivity of homology searching as well as the accuracy of the alignments that are necessary to build homology models based on existing structures. Broadly speaking, one can divide such efforts into two classes.

The first class comprises methods that seek to enhance the effectiveness of pure sequence-based queries by taking advantage of multiple-sequence alignment information in the form of profiles from particular protein families. The use of "signatures" -- short, amino acid sequence motifs --, judged to be characteristic of a family of related proteins, in the PROSITE and BLOCKS databases, represent such an effort (Henikoff, *et al*), as does the use of profile analysis in the recently developed PSI-BLAST (Altschul *et al*). Recently, Hidden Markov Models (HMMs) have been used to represent the information contained in multiple-sequence alignments. For example, the Pfam database (Sonnhammer *et al*) contains HMMs that represent various protein domains. It has been observed that the success of HMMs as homology detection devices is highly dependent on the quality of the initial alignments used to generate them (Henikoff *et al*).

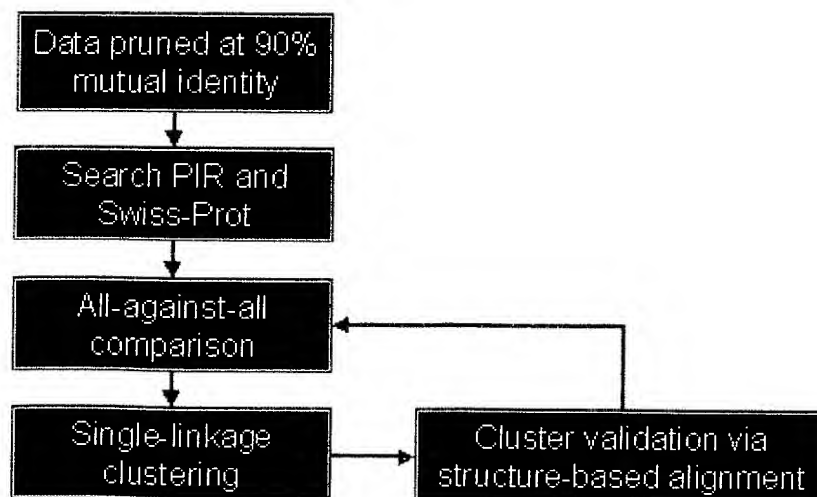
The second class comprises efforts to develop "fold recognition" algorithms that make direct use of information about structural features in the targets. The term "threading" is often used to refer collectively to such methods, though it was first used to refer to the use of mean-force potentials of pairwise residue distances, as derived from databases of protein structures (Sippl). Alternatively, several related methods have been developed in the last few years which use statistical preferences of amino acids for certain structural environments to align an amino acid sequence to an environment string using dynamic programming. Such methods may also modify gap penalties to take advantage of knowledge of where insertions or deletions could realistically be made in a target structure. While some successes have been reported in identifying very remote homologues, a recent study (Jaroszewski *et al*) observed that structure-only based searching methods did not perform as well over a set of several fold recognition benchmarks as sequence-only based methods. The best results were observed using hybrid schemes with mixed sequence/structure scoring matrices.

The goal of the MOE project described in this paper is to improve homology identification and modeling by taking advantage of both the information implicit in multiple-sequence alignments, and the structural information available from experimental models. To do this, the structures in the Protein Databank were clustered into sets of proteins with related sequences and similar structures, using sequence and structure-based alignment methods to validate the hypothesized clusters and to guarantee that the alignments - which will be used for homology searching and modeling - accurately represent the structurally conserved cores of protein families.

This paper introduces the protein family database distributed in version 1998.10 of Chemical Computing Group's Molecular Operating Environment, describes the automated protocol used to build the database, and presents the test results which demonstrate improvements in the sensitivity of homology searches.

Methods and Materials

The raw material of this study included the contents of the September 1998 release of the Brookhaven Protein Databank (Bernstein *et al*), as well as the contents of two public domain sequence databases: Swiss-Prot version 37 (Bairoch and Apweiler) and the Protein Identification Resource (PIR) version 57.0 (Barker *et al*). Only high resolution X-ray models from the PDB were used (3.0 Angstroms or higher); chains that were shorter than 25 residues or contained internal chain breaks or non-standard amino acids were excluded. The following flowchart summarizes the overall clustering protocol that was applied to this data:



The first step was to perform all-against-all sequence alignments between all the unique protein sequences in the PDB that

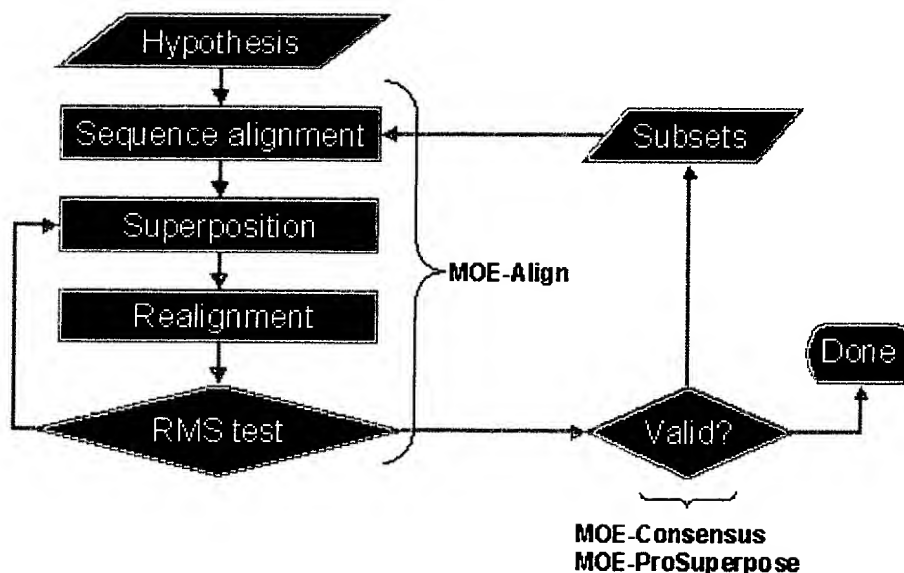
met the criteria described above. Based on the results of the sequence comparisons, links were created between any two chains whenever 90% of the residues of the longer chain were aligned against identical amino acids in the shorter chain. Using these links, the chains were clustered on a *single-linkage* basis - that is, all linked chains were put into the same cluster. Each cluster was then represented by one chain in the dataset that was submitted to the full iterative clustering procedure.

These chains were used as queries for sequence similarity searches of the PIR and Swiss-Prot databases. The criteria used to return matches from these searches were highly conservative, as there would be no structural information to confirm any hypothesized homologies. The purpose of collecting this extra sequence data prior to embarking on the iterative clustering was primarily to improve alignment quality though it was also possible that additional links between PDB chains might be identified that would otherwise have been missed.

Once the structure data had been reduced to a (relatively) non-redundant subset, and the PIR and Swiss-Prot data had been searched, the iterative phase of the procedure was entered. Each iteration began with a set of clusters. The first iteration began with a set of clusters containing one PDB entry each, and possibly augmented by sequences recruited from PIR or Swiss-Prot.

Using MOE-Align, with the sequence-only option enabled, each cluster was aligned against each other cluster, and a Z-score recorded. Z-scores were calculated by comparing raw alignment scores to a random distribution generated by random permutations of sequences in one or the other of the clusters being aligned. Then, single-linkage clustering was performed, where a link was considered to exist between two clusters if the Z-score was high enough.

This procedure created a set of hypothesized clusters, which were submitted to a validation protocol, which, roughly speaking, sought to establish whether or not all of the chains in the proposed cluster were "sufficiently superposable". The validation procedure, which involves the use of the MOE tools MOE-Align, MOE-Consensus and MOE-ProSuperpose, is summarized by the following chart:



MOE-Align was used to applied sequence and structure based alignment to the hypothesized cluster. This procedure is discussed in some detail in an earlier JCCG feature ([Multiple Sequence and Structure Alignment in MOE](#)), but a brief summary is as follows:

1. Multiple sequence alignment, using tree-based build-up and randomized iterative refinement.
2. Global-multi-body superposition of the alpha-Carbon traces

3. Re-alignment using dynamic programming, where the scoring matrix was derived from the 3D positions of the alpha-Carbons after superposition.
4. Re-superpose using the new alignment. If the RMSD has improved, go to step 3. Otherwise terminate.

After the alignment stage, the hypothesized cluster would be tested for acceptance. In view of the express purpose of this database - which is to create clusters from which accurate alignments and models can be created - the acceptance criteria was as follows :

The maximal set of alignment positions such that the worst pairwise RMSD, over these positions was less than 3.0 Angstroms was determined. If this set of alignment positions spanned a sufficient percentage of length of each chain (75%), then the cluster was accepted. otherwise it was rejected. If the cluster was rejected, then all subsets with more than two members would be tested in the same fashion.

The iterative clustering procedure was terminated when an iteration failed to produce any new clusters.

Results and Discussion

The 2300 chains were distributed into 755 families in the final database. 500 clusters contained only one PDB entry; there were 83 entries with more than 3 chains. By way of comparison, the SCOP database (Murzin *et al*) release 1.37, based on the October 1997 version of the Brookhaven PDB, distributed the proteins into slightly more than 800 "families" grouped into over 600 "superfamilies", where families were considered possess clear evolutionary relationships, and superfamilies "probable" evolutionary relationship. The clusters in MOE's family database generally corresponded to SCOP's families, though some spanned part or all of a superfamily. Some SCOP families were distributed in the MOE database into more than one cluster if the members were not adequately superposable. For example, the kinases were distributed into three families, with twitchin (PDB entry 1KOB) isolated by virtue of divergence form the othe kinases in the C-term region, and the insulin dependent kinases, which superpose to an RMSD of about 4.7 Angstroms to the other kinases, also put in their own cluster. There were other instances where SCOP families were merged - for example, the trypsins and serine proteases can by globally superposed to within 1.4 Angstroms RMSD (see [Protein Analysis in MOE: The Serine Proteases](#)).

As one of the express purposes of building this database was to improve the efficiency of remote homology detection, the following test was performed. The clusters were examined for instances of chains with less then 25% pairwise sequence identity to at least two other members in its cluster. Each such chain was then extracted from its cluster, and any chains of higher than 25% percent identity to it were discarded, and the remaining chains were re-aligned. Then, the extracted chain was aligned - using sequence information only - against each remaining member of the cluster, and against the cluster as a whole, and the maximum Z-scores achieved in the paiwise alignment was compared to the Z-score calculated against the cluster. For comparison's sake, same measurements were then made using 25% and 50% percentage identity thresholds. The results are summarized in the following table.

Maximum %id	Pairwise Z-score	Cluster Z-score	Difference
< 25	7	10.5	3
25 - 50	13	17	4
>50	32	26	-6

It is worth noting that the increase in the average Z-score among the more remote homologues resulted in lifting the strength of the sequence homology signal well out of the "noise" range. The table below contains some example of the boost in Z-scores that were observed.

Family	PDB entry	% identity	Pairwise Z-score	Cluster Z-score
Globin	1ITH.A	22	8.3	13.5
Lectin	1LCL	25	9.7	10.9
Isocitrate dehydrogenase	1IDC	23.6	12.3	14.7

While comprehensive comparison to other standard searching methods has yet been made, there are various examples in the database of clusters which include remote homologues which appear not be detected by PSI-BLAST or Pfam. For example, consider the ferredoxin oxireductase represented in the PDB by accession number 1A8P.



In MOE's family database, 1A8P was clustered with a set of five other protein chains. The percentage identities and pairwise RMSD values are shown in the tables below.

```

1A8P  NADPH\ :FERREDOXIN OXIDOREDUCTASE
1CNF  OXIDOREDUCTASE (NITROGENOUS ACCEPTOR)
1NDH  ELECTRON TRANSPORT (FLAVO PROTEIN)
1FNB  OXIDOREDUCTASE (NADP+(A) , FERREDOXIN(A) )
1QUE  FERREDOXIN--NADP+ REDUCTASE

```

		1A8P	1CNF	1NDH	1FNB	1QUE	1FDR
1	1A8P	: 100.0	15.4	11.9	13.2	12.2	30.7
2	1CNF	: 15.6	100.0	35.6	10.8	12.2	13.9
3	1NDH	: 12.5	36.9	100.0	12.5	12.9	16.8
4	1FNB	: 15.2	12.3	13.7	100.0	47.9	16.8
5	1QUE	: 14.4	14.2	14.4	49.0	100.0	15.2
6	1FDR	: 29.2	13.1	15.2	13.9	12.2	100.0

Pairwise RMSD

- lower triangle is pairwise superposition RMSD
- upper triangle difference between pairwise and global RMSD

		1A8P	1CNF	1NDH	1FNB	1QUE	1FDR
1	1A8P	: 0.000	0.000	0.000	0.000	0.000	0.000
2	1CNF	: 2.896	0.000	0.000	0.000	0.000	0.000
3	1NDH	: 2.934	1.570	0.000	0.000	0.000	0.000
4	1FNB	: 2.733	2.975	3.139	0.000	0.000	0.000
5	1QUE	: 2.857	3.039	3.193	0.818	0.000	0.000
6	1FDR	: 1.724	2.575	2.637	2.714	2.883	0.000

pro_Superpose: global RMSD = 2.660

These proteins all contain two globular domains: an oxioeductase FAD/NAD binding domain at the C-term, and a cytochrome reductase domain at the N-term. When 1A8P was extracted and submitted as query to PSI-BLAST, only 1FDR was picked up above the default significance threshold. Pfam version 3.3 (December, 1998) reported homology only to the FAD/NAD binding domain, and not to the N-term domain, despite the fact that models of both domains were in the database.

When the sequence of target 62 from the recently concluded structure prediction competition CASP3 was used as a query, this family was reported as globally homologous with very large Z-score (over 12). Again, PSI-BLAST and Pfam reported homologies only to the C-term. A model of target 62 was submitted to CASP3 based in this family.

The other models submitted to CASP3 based on alignments to this database were targets 55, 69 and 82. The results will be discussed when they become publicly available. With the exception of target 69, the percentage identities of these sequences to the templates found in the database were on the order of 20%.

References

- Altschul, S.F. *et al* Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389-3402 (1997)
- Barker, W. *et al* The PIR-International Sequence Database *Nucleic Acids Research* **26**:27-32 (1998)

Bairoch, A. and Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Research* **26**:38-42 (1998)

Bernstein F.C, Koetzle, T.F, Williams, G.J.B., Meyer, E.F., Brice M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. The Protein Data Bank: a computer based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**:535-542 (1977)

Henikoff, S., Pietrokovski, S. and Henikoff, J. Superior Performance in protein homology detection with the Blocks Database servers *Nucleic Acids Research* **26**:309-312 (1998)

Jaroszewski, L., Rychlewski L., Zhang, B., and Godzik, A. Fold Prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Science* **7**:1431-1440 (1998)

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **247**:436-540 (1995)

Needleman and Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, **48**:443-453 (1970)

Sippl, M.J. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology* **5**:229-235 (1995)

Sonnhammer, E. *et al*, Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* **26**:320-322 (1998)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER: _____**

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.